2004－ICS－138　(25)
2004／12／5

Semi-supervised sentence classification for MEDLINE documents

TAKAHIKO ITO,† MASASHI SHIMBO,† TAKAHIRO YAMASAKI†,††
and YUJI MATSUMOTO†

We address the task of sentence classification in Medline abstracts, in which sentences must be classified into their structural roles such as background, objective, methods, experimental results, and conclusions. With a plenty of labeled data, supervised learning would be able to accurately infer the structural roles of each sentence in the abstracts. However, it is not practical to assume abundant training data as they are expensive to construct. We therefore apply semi-supervised learning to this sentence classification task to remedy the lack of training data. Experimental results show that semi-supervised learning outperform pure supervised learning, when only a small amount of correctly labeled sentences are available.

1. Introduction

With the rapid increase in the volume of scientific literature, there have been growing demands for systems with which researchers can find relevant scientific literature with less effort. Online literature retrieval services, including PubMed[12] and CiteSeer[8], are gaining popularity, as they permit users to access and search through a large database of abstracts or full texts of scientific papers.

PubMed helps retrieval of medical and biological papers by means of keyword-search in the Medline abstracts[10]. It also provides a number of ways to filter search results. For example, it enables search on a specific field such as titles and journals, and filtering with publication dates. Because most abstracted papers are given peer-reviewed annotation of topics and keywords chosen from a controlled vocabulary, these can also be used to restrict search. All these facilities, however, rely on information external to the contents of abstract text. This paper exploits information inherent in abstracts to make the retrieval process more goal-oriented.

The information we exploit is the structure underlying abstract text. Our system allows search to be performed on a limited sections of an abstract, where 'sections' are determined in accordance with the structural roles of individual sentences, such as Background, Objectives, (experimental) Methods, (experimental) Results, and Conclusions.

To see the advantage of such a system, notice that some of these 'sections' are more relevant to the goals of users compared with the rest. Hence, by specifying these relevant sections as the target of search, the literature retrieval process can be made more goal-oriented. For example, a clinician, trying to find whether an effect of a chemical substance on a disease is known or not, can ask the search engine for passages containing both the names of the substance and the disease, but only from the Results and Conclusions sections. Such restriction is not easily attainable by simply adding extra query terms. Moreover, it is not always evident to users what extra query terms are effective for narrowing down search results. Specifying target sections should be helpful in this case as well.

A major challenge in building such a system is how to infer sectioning of abstracts. This process must be (semi-)automated, because it is too expensive to manually section abstracts due to the size of Medline. In our previous work[14),17)], we reported preliminary results on the application of supervised text classification techniques to this task, using the 'structured abstracts' as labeled data.

Structured abstracts[1)] are abstracts in which sections are explicitly marked with headings. These headings are customarily written in all upper-case letters, and are therefore easily identifiable. Typical headings include BACKGROUND, OBJECTIVE, METHOD, RESULTS, and CONCLUSION. As seen from this list, headings indicate the structural role (i.e., class) of the sentences following them.

In most research fields other than biology and medicine, however, abstracts are not structured at all. Even in structured abstracts, the labels translated from section headings do not always reflect the actual structural roles of the sentences contained in the corresponding sections. We pointed out that

† 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
†† 現沖電器 (株)
Currently with OKI Co.

the sentences with BACKGROUND heading frequently contain the objective of the work, which should rather be labeled as OBJECTIVES class[15]. The main reason for such discrepancy should be that most journals do not give a concrete principle on how and what section headings should be presented in an abstract, and hence these are determined by diverse authors each with his/her own criterion.

This observation motivates us to formulate the task of sectioning abstracts not in a pure supervised learning setting, but in a semi-supervised learning setting, in which a small amount of manually labeled data is combined with a larger amount of unlabeled data. In this formulation, sentences in structured abstracts are treated as unlabeled data. The heading of a sentence is not a class label anymore, but is a feature that merely suggests a probable label for the sentence. Note here that the sentences we plan to classify generally do not have headings at all, so this feature should have utility only in semi-supervised setting (i.e., in the training phase).

As preliminary work, we apply a semi-supervised learning technique, namely Transductive SVMs[5], to sentence labeling tasks using manually labeled data and unlabeled data.

Note that this paper does not use structured abstracts as labeled examples. Instead we use a small amount of human-annotated unstructured abstracts. Although we plan to use structured abstracts eventually, they are not used here to avoid the noise they may introduce. The main focus of this paper is therefore to evaluate whether semi-supervised setting is potentially worth attempting for this task.

This paper proceeds as follows. In Section 2, we introduce Support Vector Machines[16] and its transductive extension due to Joachims[5]. We then present an experimental result to examine the performances of transductive formulation. Finally, we conclude and describe future work.

2. Classifier design

In our previous work[14),17)], we classified the sentences in Medline into five classes: BACKGROUND, OBJECTIVE, METHOD, RESULTS, and CONCLUSION. The choice of these classes is based on the frequency of individual headings in the structured abstracts contained in Medline 2002 (Table 1). This paper also use these five classes.

This section first describes the application of standard inductive Support Vector Machines[3),16)]. to our sentence classification tasks, followed by the introduction of the framework of transductive SVMs. The features used in our experiments will be described as well.

Table 1 Frequency of individual sections in structured abstracts from Medline 2002.

Sections	# of abstracts	# of sentences
CONCLUSION(S)	352,153	246,607
RESULTS	324,479	1,378,785
METHODS	209,910	540,415
BACKGROUND	120,877	264,589
OBJECTIVE	165,972	166,890
⋮	⋮	⋮

2.1 Support Vector Machines

The application of standard, purely inductive, Support Vector Machines to the present task follows our previous work[15]. We first construct a soft-margin SVM classifier for each of the five classes, BACKGROUND, OBJECTIVE, METHOD, RESULT, and CONCLUSION, using the one-versus-rest configuration, in other words, by taking the sentences in the target class as the positive examples, and the rest of the classes as the negative examples.

Since SVM is a binary classifier while our task is a multi-class classification involving five classes, we combine the results of these classifiers to determine a single class label assigned to a given sentence, as follows.

For each class index $i = 1,\ldots 5$, let $f_i(\cdot)$ be the decision function of the SVM representing the i-th class, i.e., $f_i(x)$ gives the signed distance of an example x from the optimal hyperplane in the feature space. ☆. Given a test example x, the class label assigned to x is $i = \arg\max_i f_i(x)$. Thus, when there are classes i that satisfies $f_i(x) > 0$, the class whose separating hyperplane is farthest apart from x is chosen from these classes; If there is no classes with $f_i(x) > 0$, the class with separating hyperplane nearest to x is assigned to x.

2.2 Transductive SVMs

The notion of Transductive SVMs was first introduced by Vapnik[16]. Transductive SVMs use labeled data as well as unlabeled data to construct "optimal" separating hyperplanes. Since SVMs deal with binary classification, each labeled exam-

☆ SVMs minimize the normal vector w of separating hyperplanes under constraint $wx+b = \pm 1$ for every support vector x, with b being a bias term. It follows that given a separating hyperplane with normal w and bias b, we have the signed distance $f(x) = (wx+b)/|w|$ of point x from the hyperplane.

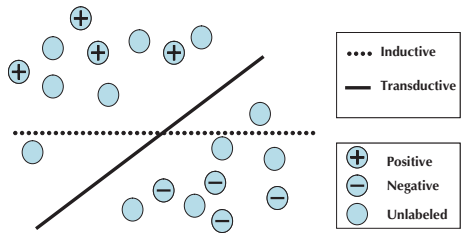

Fig. 1 Hyperplane with inductive and transductive SVM: Transductive SVMs change the hyperplanes of inductive SVMs to maximum margins based on not only labeled examples but unlabeled examples

ple is associated with a label of either "positive" or "negative." On the other hand, we do not know the labels of unlabeled examples. Transductive SVMs first infer the labels of unlabeled examples from their features, and then compute the maximum margin between "positive"and "negative"examples, including those unlabeled examples with inferred labels. Construction of a Transductive SVM reduces to an optimization problem similar to that of an inductive SVM. Unfortunately, the optimization for Transductive SVMs turns out to be much harder to solve than the inductive SVMs. To get around this problem, Joachims[5] proposed to approximate the solution by iterative application of SVMs, gradually changing the population of positive and negative labels assigned to unlabeled data. Henceforth, by Transductive SVMs, we mean Joachims's approximation algorithm.

In our classification task in which each sentence in an abstract is classified into classes representing structural roles, manually sectioning abstracts is often expensive since the contents of abstracts are highly specialized. To cope with the limited number of sectioned abstracts, we apply to our task Joachim's Transductive SVMs, which are known to often perform better than inductive SVMs when the amount of labeled data is small. Therefore, to improve the accuracy of sentence classifiers, we use not only a small amount of abstract sentences manually annotated with their structural roles, but also a large amount of sentences in which no roles are identified.

2.3 Feature representation

We use the so-called "bag-of-words" to represent features for learning and classification. Although non-contiguous sequential word patterns may be used as well[14),17)], we only use words here because the speed of feature construction would be impractical due to the large number of abstracts to be processed.

In addition, we incorporate additional features that represent the context in which a given sentence occurs. Since we are interested in labeling a *series* of sentences, incorporating contextual information into the feature set is expected to improve classification performance.re These features would allow us to capture the trends such as (1) it is unlikely that experimental results (RESULT) are presented before the description of experimental design (METHOD), and (2) the sentences of the same class (i.e., with the same structural roles) have high probability of occurring consecutively; we would not expect authors to interleave sentences describing experimental results (RESULT) with those in CONCLUSION and OBJECTIVE classes.

In our previous work[15)], the effectiveness of context features was demonstrated in the experiments in a pure inductive setting, in which structured abstracts were used as labeled data. This paper examines the effectiveness of context features under semi-supervised learning settings (see Section 3.3).

3. Experiments

This section reports the results of the experiments we conducted to examine the performance of sentence classifiers.

3.1 Experimental Setting

In this experiment, we used 34703 manually labeled sentences from 4185 abstracts randomly sampled from the set of the structured and unstructured abstracts from Medline 2002. Five classes (labels) were adopted for each sentence in collected abstracts, BACKGROUND, OBJECTIVE, METHOD(S), RESULTS, and CONCLUSION(S) as our previous work[14),17)].

All of following experiments were done with inductive and transductive SVMs using linear kernels and varying soft-margin parameter C of SVMs among $\{0.01, 0.1, 1, 10, 100\}$. 5-fold cross validation was employed and the results are therefore based on the averages of the five trials over different partitions of training and test sets.

We used only a fraction of training data in 5-fold cross validation for standard inductive SVMs. In transductive SVMs, we also added the remaining training data that are not selected as labeled data removing their labels.

3.2 Inductive and transductive SVMs

Inductive Support Vector Machines (SVMs) and transductive SVMs were applied to discriminate

one class from the others. No filtering of words are performed, to avoid null feature vectors. In addition, we use context feature representations. To put it concretely, word features of the previous and next sentences were applied.

The results of the experiments are shown in Tables 2 and 3. The performance of transductive and inductive SVMs were commonly evaluated by three measures, precision, recall and F-measure[9]. The experiments were done varying the size of labeled data used in SVMs among $\{100, 1000\}$ and using words as features.

The transductive SVMs outperformed inductive SVMs in F-measure, not only when the number of labeled examples were small (100), but also there were a larger number of labeled data (1000). Although the transductive SVMs were inferior in precision, they significantly outperform inductive SVMs in recall in most of classification.

3.3 Context Features

Next, we examined the performance of context features in transductive SVMs. As context features, words in the previous and next sentence were employed. Table 4 only shows the average F-measure. Columns "with context" and "without context" respectively display the performance (F-measure) of inductive and transductive SVMs with and without context features.

Inductive and transductive SVMs with context features outperformed those without contexual features in most classes. For example, in classifying BACKGRAOUND, transductive SVMs with context features obtain about 66 point F-measure when there are 1000 labeled data, which is an improvement of 13 points over the classifier without context features.

3.4 System performance

We also examined the performance of this system combining classifiers for each label (see Section 2.1). In this experiments, transductive and inductive SVMs were compared by accuracy, and in addition, we checked performance of the systems with context features and without them. Table 5 shows the performance of the systems that classify sentence in abstracts into five class.

It is true that the transductive SVMs with 100 labeled examples were inferior in F-measure when they did not use context features, but, with 1000 labeled example, they outperformed inductive SVMs, whether or not context features were applied. For example, transductive SVMs with 1000 labeled data was 66 point, an 8 points improvement over inductive SVM.

4. Conclusions

4.1 Summary

We have proposed to apply semi-supervised learning (transductive SVMs) to infer structural roles of the sentences in Medline abstracts. Transductive SVMs achieved higher performance than inductive SVMs not only in class-wise classification accuracy, but also in the overall multi-class classification accuracy after the classification results for individual classes were combined.

Experimental results confirmed that Transductive SVMs in this task demonstrates higher performance than inductive SVMs especially when there is only a small amount of labeled data.

We have also demonstrated the effectiveness of contextual features in this task. With a small dataset, we obtained an approximately 66% classification performance in our sentence classification systems by combining transductive SVMs and contextual features.

4.2 Future work

In this work, only one semi-supervised classification method (Transductive SVMs[5]) is applied to sentence classification. However, many types of semi-supervised classification have been newly proposed[2),6),11)]. Among them, we plan to select the semi-supervised algorithms most suitable for our task.

Futhermore, to improve classification performance, we will incorporate structured abstracts into semi-supervised framework. In this setting, each section heading in structured abstracts is not an indicator of the section label but merely a feature which is not always present. The sentences in structured abstracts are treated as an unlabeled examples as a result.

We also plan to incorporate a re-ranking procedure of label sequences based on the overall consistency of sequences. Here, by 'consistency' we mean the degree to which constraints on label sequences are fulfilled, where constraints include rules such as conclusions never appear in the beginning of an abstract, and the sections hardly occur interleaved with each other. Although some of these constraints can be captured through the context features of Section 2.3, a more elaborate mechanism is desirable; the context features only affect local decisions performed sequentially, but does not

Table 2 Performance of classification (# of labeled example = 100)

	Inductive SVMs			Transductive SVMs		
	precision	recall	F-measure	precision	recall	F-measure
BACKGROUND	62.6	18.3	26.8	49.4	47.1	47.1
OBJECTIVE	34.3	3.5	5.7	20.9	29.4	23.7
METHOD(S)	36.9	5.2	8.9	33.6	29.9	31.3
RESULTS	64.9	48.7	55.5	61.8	59.0	60.2
CONCLUSIONS	71.9	22.4	30.9	58.9	49.3	53.0

Table 3 Performance of classification (# of labeled example = 1000)

	Inductive SVMs			Transductive SVMs		
	precision	recall	F-measure	precision	recall	F-measure
BACKGROUND	71.1	54.9	61.8	66.8	65.4	66.0
OBJECTIVE	66.6	25.1	35.3	54.1	49.1	51.0
METHOD(S)	58.6	36.0	44.4	50.7	48.2	49.3
RESULTS	73.7	68.1	70.8	72.8	73.0	72.9
CONCLUSIONS	74.5	56.0	63.9	68.9	66.7	67.7

Table 4 Performance of context features (F-measure)

	# of labeled example = 100		# of labeled example = 1000	
	with context	without context	with context	without context
BACKGROUND	47.1	33.6	66.0	53.1
OBJECTIVE	23.7	25.2	51.0	51.1
METHOD(S)	31.3	35.9	49.3	50.1
RESULTS	60.2	58.8	72.9	71.6
CONCLUSIONS	53.0	29.4	67.7	52.7

Table 5 Performance of sentence classification system (accuracy)

	# of labeled example = 100		# of labeled example = 1000	
	Inductive SVMs	Tranductive SVMs	Inductive SVMs	Transductive SVMs
with context	0.46	0.47	0.62	0.66
without context	0.45	0.36	0.47	0.54

take into account the 'global' consistency of a label sequence as a whole.

The similar lines of research[7),13)] have been reported recently in machine learning and natural language processing communities, in which the sequence of classification results is optimized over all possible sequences.

Devising features that reflect cohesion or coherence between sentences[4)] is another interesting topic to pursue.

References

1) Ad Hoc Working Group for Critical Appraisal of Medical Literature. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine*, 106(4):598–604, 1987.
2) A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *International Conference on Machine Learning*, 2001.
3) C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.
4) M. A. K. Halliday and R. Hasan. *Cohesion in English*. Longman, London, 1976.
5) T. Joachims. Transductive inference for text classification using support vector machines. In I. Bratko and S. Dzeroski eds., *Proceedings of ICML-99, 16th International Conference on Machine Learning*, pp. 200–209, Bled, SL, 1999. Morgan Kaufmann Publishers, San Francisco, US.
6) T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of 20th International Conference on Machine Learning*, pp. 290–297. Morgan Kaufmann Publishers, San Francisco, US, 2003.
7) J.Lafferty, A.McCallum, and F.Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, pp. 282–289. Morgan Kaufmann, 2001.
8) S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.
9) C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press,

fourth edition, 1999.
10) MEDLINE. http://www.nlm.nih.gov/databases/databases_medline.html, 2002–2003. U.S. National Library of Medicine.
11) K. Nigam, A. McCallum, and S. T. T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2/3):103–134, 2000.
12) PubMed. http://www.ncbi.nlm.nih.gov/PubMed/, 2003. U.S. National Library of Medicine.
13) F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proceedings of the Human Language Technology Conference North American Chapter of Association for Computational Linguistics (HLT-NAACL 2003)*, pp. 213–220, Edmonton, Alberta, Canada, 2003. Association for Computational Linguistics.
14) M. Shimbo, T. Yamasaki, and Y. Matsumoto. Automatic classification of sentences in the MEDLINE abstracts. In *Proceedings of the 6th Sanken (ISIR) International Symposium*, pp. 135–138, Suita, Osaka, Japan, 2003.
15) M. Shimbo, T. Yamasaki, and Y. Matsumoto. Automatic classification of sentences using sequential patterns. In *Proceedings of the Second International Workshop on Active Mining*, pp. 32–41, Maebashi, Japan, 2003.
16) V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.
17) T. Yamasaki, M. Shimbo, and Y. Matsumoto. Automatic classification of sentences using sequential patterns. Technical Report of IEICE AI2002-83, The Institute of Electronics, Information and Communication Engineers, 2003. In Japanese.